

周报（2013.10.21-2013.10.27）

本周工作：

1. 完全 feature selection 各种方法的小综述，具体内容见后面。
2. 完成翻译内容的校对
3. 准备主题报告

下周工作：

1. 调研 local svm

目前存在的各种 feature selection 主要包括以下几个方面：

1. 变量排序（variable ranking）

定义在单个变量上，独立于其他的上下文，相关性分析方法（Correlation method）就属于这类方法。变量排序简单、扩展性好，通常也作为一种过滤方法，是一种预处理手段，与预测指标选取无关。不过一般只能判断线性关系。

对于一个实例 $\{x_k, y_k\}$ ($k=1,2,3,\dots,m$)，有 n 个输入变量 $x_{k,i}$ ($i=1,2,3,\dots,n$)，一个输出变量 y_k 。从 $x_{k,i}$ 和 y_k 构建一个评分函数（scoring function） $S(i)$ 。通常得分越高，认为该变量越重要，按所有变量的 $S(i)$ 降序排列。可采用的方法有相关系数和信息理论准则。

相关系数

把向量 x 看成符合某种潜在未知分布的实例化， X_i 表示 x 中的第 i 个量，同样 Y 表示输出 y 的随机实例化。

对于连续输出 y 的预测，Pearson correlation coefficient 定义为：

$$R(i) = \frac{\text{cov}(X_i, Y)}{\sqrt{\text{var}(X_i)\text{var}(Y)}}$$

cov 表示两个变量间的协方差，var 表示单个变量的方差

$R(i)$ 也可用于线性回归和二元分类。

$R(i)$ 这种相关性准则的一个缺点是只能判断线性关系。一种改进方法是选择一个非线性的目标值，比如平方、开方、取对数、求逆

$R(i)^2$ 可根据单个变量的预测能力进行分类。比如，对一个变量的取值范围设定一个阈值。评价分类的能力除了正确率，还可选用 false positive classification rate 和 false negative classification rate. 如果多个变量都有很好的分类正确率，可以选择其他的统计值，比如类之间的 margin（svm 中采用）

信息理论

基于每个变量和目标之间的相互信息：

$$I(i) = \int_{x_i} \int_y p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dx dy$$

其中 $p(x_i)$ 和 $p(y)$ 是 x_i 和 y 的概率密度， $p(x_i, y)$ 是联合概率密度。 $I(i)$ 反映了 x_i 和 y 之间的依赖关系。这种方法的缺点之一是 $p(x_i)$ 、 $p(y)$ 和 $p(x_i, y)$ 一般未知，并且很难从数据中估计。对于离散型或类别型数据，这个公式可以简化为

$$I(i) = \sum_{x_i} \sum_y P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)}$$

不过存在计算量大的问题。

对于连续变量信息理论这种方法最困难，可以把变量离散化或者用非参数方法估计密度
关于变量相关性、冗余的一些重要的观点：

- 1) 可以通过添加一些看似冗余的变量达到去除噪声和更好的分类结果
 - 2) 完全相关的变量只有加上这些变量没有得到任何额外信息，才认为这些变量是冗余的（也就是看似完全相关的变量不一定真的冗余）
 - 3) 一个变量本身可能没什么作用，与其他变量一起考虑时可能很有意义（XOR 问题）
- 这些思想在变量排序的时候要考虑，同时也说明了选择一个变量的子集和不是单独的每个变量很重要。

2. 变量子集选择

Feature subset selection 一般包括四个步骤：产生子集(subset generation)、subset evaluation, stopping criterion, result validation。首先根据某种搜索策略产生候选子集。对每个产生的候选子集根据 evaluation criteria 进行评估，并与之前的产生的最好的 subset 比较。如果当前的 subset 更优，则替换之前最优的 subset。Subset generation 和 subset evaluation 的过程一直重复直到达到某种 stopping criteria。最后，产生的 best feature subset 用已知的先验知识或者不同的数据集进行测试。通常会选择一个 evaluation criteria 而不是该子集的分类正确率（或者识别正确率）来评估一个 feature subset 的性能，因为往往训练数据集会很大，计算分类正确率会比较耗时。

选择变量子集主要包括三种方法：wrappers, filters, embedded method.

Wrapper 方法需要一种预先设定的数据挖掘算法作为黑盒，用该算法的性能作为评价准则。因此 wrapper 方法是为了提高特定数据挖掘算法的性能而进行 feature selection。往往计算量更大。filter 是作为预处理来选择变量子集，与预测指标、挖掘算法的选取无关。相对比较简单，计算量比较小。Embedded method 指在训练的过程中进行变量子集的选择。Wrapper 一般更通用、简单，不过 embedded method 在训练中选择变量，某些方面会更有效，比如不需要训练集，更好地利用了数据；不需要重新训练预测指标，可以更快地得到结果。决策树方法就是一种 embedded method。

在变量子集空间进行搜索的策略一般包括完全搜索（complete search）、顺序搜索（sequential search）和随机搜索（random search）。Complete search 一般可以获得全局最优的子集，不过时间耗费多。穷举搜索就是一种 complete search 的策略。不过完全搜索不一定意味着要穷举所有子集，可以采用一些启发式搜索策略，比如 branch and bound、beam search。Sequential search 可能会获得局部最优值。常采用的一些贪心爬山算法的变种包括：sequential forward selection、sequential backward elimination 或者 bi-directional selection。Forward selection, backward elimination. forward selection 指初始的 subset 为空，后面逐渐把变量添加到集合中，backward elimination 方法的初始集合包含所有变量，然后一步步把集合中的某些变量去掉。一般后向选择会比前向选择更慢，但往往可以得到更稳定的 feature subset。前向选择通常更简单，速度相对更快一些。为了提高速度，对其的一种改进方法是每次往集合中添加多个变量或者从集合中删除多个变量。通常 sequential search 的搜索效率比 complete search 高，搜索空间一般只有 $O(n^2)$ 或者更小。随机搜索（random search）一般包括两种方法，一种是在 sequential search 中引入随机性，比如 random-start hill-climbing 或者 simulated annealing，另一种是完全随机地产生下一个候选子集，比如 Las Vegas algorithm。随机搜索（random search）的这些方法引入随机性都是为了避免陷入局部最优。

在所有的子集选择方法中，Subset evaluation 中一般采用的评价准则根据其对于数据挖掘算法的依赖性分为两种：独立性准则（independent criteria）、依赖性准则（dependent criteria）。通常独立性准则用于 filter 方法中，通过探究训练数据的本质特征来评价特征或特征集合，不涉及任何数据挖掘算法。常用的独立性准则包括距离测度（distance measure）、信息测度（information measure）、依赖性测度（dependency measure）、一致性测度（consistency measure）。Distance measure 一般也指分离性测度或者可判别性测度。在条件概率中，如果一个特征 X 相比于另一个特征 Y 能更好地对数据进行区分，X 就优于 Y。information measure 一般用来确定从一个特征得到的信息增益。如果从一个特征 X 中获得的信息比从特征 Y 中获得的信息多，就认为特征 X 更好。Dependency measure 一般指关联性测度（correlation measure）或者相似性测度（similarity measure）。在分类问题中，如果一个特征 X 比另一个特征 Y 与类别标签更相关，就认为 X 更优。Consistency measure 指试图找到一个最小的特征集合可能与原始所有特征具有一致的分类性能。Dependency criteria 一般用于 wrapper 中，在特征选取中需要预定义一种数据挖掘算法，这种准则通常会选择更适合给定数据挖掘算法的特征，因此往往性能比较好，不过计算量会更大，并且可能不适合所有的数据挖掘算法。在聚类问题中，wrapper 方法通过用特征子集进行聚类的聚类结果性能来评价特征子集的优良。用来估计聚类结果的启发性准则包括：类密集度（cluster compactness）、分散性（scatter separability）、最大似然性（maximum likelihood）等。

Embedded method

Embedded method 一般包括两大类：nested subset method 和 direct objective optimization.

Nested subset method

Nested subset method 指通过估计变量子集空间移动引起的目标函数值的变化来引导搜索，并与贪心方法结合。

s 为选择的变量数目， $J(s)$ 表示目标函数，预测目标函数的变化包括以下步骤：

1) 计算有限差（finite difference）。 $J(s)$ 与 $J(s-1)$ 或者与 $J(s+1)$ 的差以决定是为了添加或者删除候选变量

2) 计算损失函数的二次逼近（Quadratic approximation of the cost function）。把

J 写成二阶泰勒展开式。当 J 达到最优时，忽略一阶项，于是 $DJ_i = (1/2) \frac{\partial^2}{\partial w_i^2} (Dw_i)^2$

的变化 $Dw_i = w_i$ 指要把变量 i 去掉。

3) 目标函数敏感度计算： J 对 x_i 或者 w_i 的导数的绝对值或者平方用来计算敏感度。

这种思路包括三种方法：

方法一：有限差（finite difference）：比较简单容易计算，不需要对候选变量重新训练模型。比如线性最小二乘模型，在执行向前变量选择时每次添加均方差最小的变量。也可采用近似差：比如像核函数方法（kernel method），学习形式如 $f(x) = \sum_{k=1}^m \alpha_k K(x, x_k)$ 的机器， $K(x, x_k)$ 表示核函数，测量 x 与 x_k 之间的相似性。使 α_k 保持不变来计算 $J(s)$ 的变化。这种方法最初用于 svm。

方法二：optimum brain damage (OBD)。OBD 认为线性预测值的方法比较简单，因此采用 DJ_i 而不是 $|w_i|$ 作为剪枝标准。不过如果把目标函数 J 表示成 w_i 的二次方的形式，两者等价。

方法三：前向变量选择中利用目标函数对 w_i 变化的敏感度。这样选择的标准是

$\frac{\partial J}{\partial w_i} = \sum_{k=1}^m \frac{\partial J}{\partial \rho_k} \frac{\partial \rho_k}{\partial w_i}$ ，其中 $\rho_k = y_k f(x_k)$ 。这种方法的一个有趣的变形是把目标函数换成留

一法交叉验证的错误率。

Direct Objective Optimization

通常目标函数包括两项：拟合度 (goodness-of-fit) (需要最大化) 和变量的数目 (要最小化)。有的论文中用一个 regularization 项代替变量数目这一项，这个 regularization 项一般是用于缩小参数空间

Filters

Filters 相比于 wrapper 速度更快，一般是作为一种通用的变量选择方法，不针对某种特定的学习机器。可用于预处理来降维和避免过度拟合。一种做法是先用 wrapper 或者 embeded method 构造线性预测指标来过滤，然后对于选出来的最终变量构造一个更复杂的非线性预测指标。

3. 特征构建和降维

常用的特征构建方法有：聚类、输入变量的基础线性变化 (如 PCA/SVD, LDA)、更复杂的线性变化 (如光谱分析、傅里叶变换等)、小波变换、核卷积。特征构建一般有两个目标：更好地重构原数据或者更有效地进行预测。

聚类

最常用的聚类方法是 K-means 和层次聚类 (hierarchical clustering)

矩阵分解

Singular value decomposition (SVD) 的目标是构建原始变量线性组合的一组特征，是一种无监督方法。

4. svm 中 feature selection 的一个性质是：从 svm 中删除某一维度，分类的 margin 会变小或者不变；当最优的超平面平行于该轴时，margin 不变，否则变小。